METHOD AND APPARATUS FOR LOAD BALANCING WORK ON A NETWORK OF SERVERS BASED ON THE PROBABILITY OF BEING SERVICED WITHIN A SERVICE TIME GOAL

FIELD OF THE INVENTION

The present invention is related to a method and apparatus for load balancing work. In particular, the present invention is directed to load balancing work based on a relative probability that a server will service work within a predetermined interval.

BACKGROUND OF THE INVENTION

Call centers are systems that enable a pool of agents to serve incoming and/or outgoing calls, with the calls being distributed and connected to whichever of the agents happen to be available at the time. When no agents are free and available to handle additional calls, additional incoming calls are typically placed in a holding queue to await an available agent. It is common practice to divide the pool of agents into a plurality of groups, commonly referred to as splits, and to assign different types of calls to different splits. For example, different splits may be designated to handle calls pertaining to different client companies, or calls pertaining to different products or services of the same client company. Alternatively, the agents in different splits may have different skills, and calls requiring different ones of these skills are then directed to different ones of these splits. Each split typically has its own incoming call queue.

Furthermore, some large companies find it effective to have a plurality of call centers, each for handling calls within a different geographical area, for example, Each call center, or each split within each call center, typically has its own incoming call queue. In a multiple queue environment, it can happen that one call center or split is heavily overloaded with calls and has a full queue of calls waiting for an available agent,

while another call center or split may be only lightly overloaded and yet another call center or split may not be overloaded at all and actually may have idle agents. To alleviate such inefficiencies, some call centers have implemented a capability whereby, if the primary (preferred) split or call center for handling a particular call is heavily overloaded and its queue is overflowing with waiting calls, the call center evaluates the load of the other (backup) splits or call centers to determine if one of the other splits or call centers is less busy and consequently may be able to handle the overflow call and do so more promptly. The overflow call is then queued to the first such backup split or call center that is found, instead of being queued to the primary split or call center. Such arrangements are known by different names, one being "Look Ahead Interflow."

In order to balance work across a network of call centers, the decision as to where a call should be routed is typically made based on the estimated waiting time that a call will experience with respect to a particular switch. The objective is to find the switch within a network of switches where it is predicted that the call will be answered in the shortest period of time. In situations where an enterprise has contracted with its customers to service calls within a given period of time, sending calls to the switch with the shortest waiting time does not necessarily maximize the number of customers who are serviced within the contracting period. In particular, although doing so will generally reduce the average waiting time of calls, this is not the same as maximizing the number of calls serviced within the contracted time.

SUMMARY OF THE INVENTION

The present invention is directed to solving these and other problems and disadvantages of the prior art. Generally, according to the present invention, work (*e.g.*, a

call) is routed to a server (*e.g.*, a switch) based on the probability that the work will be serviced within a contracted time interval. In particular, the work may be routed to the server having the highest probability for servicing the work based on the relative probabilities of each server in the network to service the work within a target service time goal. In accordance with another embodiment of the present invention, work may be routed to the server identified as having a sufficient probability of servicing the work within a target service time goal. Accordingly, the present invention is capable of efficiently routing work, and does so without performing a complicated calculation of absolute probability. Instead, only the relative probabilities need to be determined.

In accordance with an embodiment of the present invention, in response to receiving a work request, the probability of servicing the work request within a target time is determined for each server in a network. The server having the greatest determined probability of servicing the work request within the target time, or having a sufficient determined probability of servicing the work request within the target time, is selected, and the work request is assigned to the selected server. In accordance with an embodiment of the present invention, the relative probability that each server will complete the work request within the target time is calculated, rather than an absolute probability, thereby reducing the computational overhead of a method or apparatus in accordance with the present invention.

In accordance with still another embodiment of the present invention, the probability of servicing the work request within a target time is determined for a server by calculating a number of opportunities to service the work request within the target time with respect to the server. If more than one server has a greatest number of

opportunities to service the work request within the target time, or if more than one server has a sufficient number of probabilities to service the work request within the target time, one of the servers may be selected by calculating an advance time metric. For instance, in accordance with an embodiment of the present invention, the server having the lowest expected wait time may be selected. In accordance with another embodiment of the present invention, the server having the lowest weighted advance time trend is selected.

In accordance with another embodiment of the present invention, a load balancing or work allocation apparatus is provided that includes a plurality of service locations. At least one service resource is associated with each of the service locations. In addition, a communication network interface is provided, operable to receive requests. A provided controller assigns the work request received at the communication network interface to the service location having the highest probability or to a service location having a sufficient probability of servicing the work request within a predetermined target time.

These and other advantages and features of the invention will become more apparent from the following description of an illustrative embodiment of the invention taken together with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a communication arrangement incorporating a system in accordance with an embodiment of the present invention;

Fig. 2 is a block diagram depicting a switch in accordance with an embodiment of the present invention;

Fig. 3 is a flow chart depicting the assignment of work based on probability in accordance with an embodiment of the present invention;

Fig. 4 is a flow chart depicting determining a probability in accordance with an embodiment of the present invention; and

Fig. 5 is a flow chart depicting the calculation of an advance time metric in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

With reference now to **Fig. 1**, a communication arrangement incorporating a system 100 in accordance with the present invention is illustrated. In general, the communication arrangement includes a device requesting service 104 interconnected to a communication network 108. The communication network 108 is in turn connected to a number of switches 112. Associated with each switch 112 are one or more resources 116, depicted in **Fig. 1** as agents. Collectively, a switch 112 and associated resources 116 comprise a service location 120. In accordance with a further embodiment of the present invention, a service location 120 may comprise a switch 112 and a subset of the associated resources 116 established or functioning as a split. For purposes of this discussion, the term "service location" is understood to include a split. Accordingly, as can be appreciated by one of skill in the art, a system 100 in accordance with the present invention may be beneficially used to allocate requests for service among splits established with respect to resources 116 associated with a single switch 112. A system 100 in accordance with the present invention may also include a control 124.

The device requesting service 104 may comprise any device in connection with which a resource 116 is desired or required. Accordingly, a device requesting service 104 may include a telephone or other communication device associated with a user, or a computing or information device associated with a user or operating autonomously.

The communication network 108 may include a public switched telephone network (PSTN), a packet data network such as a local area network, an intranet, or the Internet, or any combination of communication networks.

The switches 112, as will be described in greater detail below, may include servers, including communication servers, such as private branch exchanges or call center servers, including but not limited to automatic call distribution systems. In general, the switches 112 operate to receive requests for service from a requesting device 104 that is delivered to the switch 112 by the communication network 108. In addition, the switches 112 operate to allocate an appropriate resource 116 to service the request. In accordance with an embodiment of the present invention, a switch 112 may function to allocate requests for service to resources 116 directly associated with the switch 112, or to resources 116 associated with another switch 112. Accordingly, the functions of the optional control 124 may be incorporated into one or more of the switches 112.

The control 124 may be provided for allocating requests for service among switches 112, or among splits comprising a group of resources 116 established in connection with one or more switches 112. Furthermore, requests for service may be placed in queues established with respect to each service location 120 or split included in a system 100. A control 124 may function to calculate the probability that each switch and/or split 112 that is a candidate for servicing a request will be successful at servicing such request within a target time, as will be described in greater detail below. Alternatively, the function of the control 124 may be performed by a switch 112 incorporating such functionality. In general, the control 124 may comprise a server

computer in communication with the switches 112 either directly or through a network, such as the communication network 108.

With reference now to **Fig. 2**, a server, such as a switch 112 or a control 124, is illustrated. In general, the server 112, 124 may comprise a general purpose computer server. For example, the server 112, 124 may comprise a general purpose computer running a WINDOWS operating system. As yet another example, when implemented as a switch 112, the server may comprise a call center server, a telecommunications switch, or a private branch exchange. As shown in **Fig. 2**, a server 112, 124 may include a processor 204, memory 208, data storage 212, a first network interface 216, and optionally a second network interface 220. The various components 204-220 may be interconnected by a communication bus 224.

The processor 204 may include any processor capable of performing instructions encoded in software. In accordance with another embodiment of the present invention, the processor 204 may comprise a controller or application specific integrated circuit (ASIC) having and capable of performing instructions encoded in logic circuits. The memory 208 may be used to store programs or data, including data comprising a queue or queues, in connection with the running of programs or instructions on the processor 204. The data storage 212 may generally include storage for programs and data. For example, the data storage 212 may store operating system code 224, and various applications, including a probability function application 228 and a work distribution application 232, capable of execution by the processor 204. The first network interface 216 may be provided to interconnect the server 112, 124 to other devices either directly or over a computer or communication network, such as communication network 108. The server

112, 124 may include an additional network interface 220, for example where the server 112, 124 functions as a call center switch 112 that serves to interconnect the switch 112 to the communication network 108 and to service resources 116.

As can be appreciated by one of skill in the art, the actual implementation of a server 112, 124 may vary depending on the particular application. For example, a switch 112 that does not compute a relative probability as described herein would not require a probability function application 228. Similarly, a server comprising a control 124 would generally feature only a single network interface 216. In addition, a server 112, 124 with a processor 204 comprising a controller or other integrated device need not include memory 204 and/or data storage 212 that is separate from the processor 204.

With reference now to **Fig. 3**, a flow chart depicting the allocation of work to one of a plurality of service locations is illustrated. Initially, at step 300, a work request is received. In general, the work request may be received at a switch 112, or at a control 124. At step 304, the service location(s) 120 at which the probability of servicing the work associated with the received work request within a target time is greatest is determined. According to another embodiment of the present invention, the service location(s) 120 at which the probability of servicing the work within the target time is sufficient is determined at step 304. A sufficient probability is, according to an embodiment of the present invention, a selected number of opportunities for the work to be served within the target time. For example, three opportunities to service work within the target time may be deemed to represent a "sufficient probability" for servicing the work. The probability that is determined is not required to be an absolute probability. Accordingly, as described in greater detail below, the determination of the service

location 120 having the greatest probability for servicing the work within the target time, or the identification of a service location 120 having a sufficient probability of servicing the work within the target time, may be made from the relative probability that an eligible service location 120 will complete the work within the target time.

At step 308, a determination is made as to whether multiple service locations 120 are determined to have the greatest probability or a sufficient probability of servicing the work within the target time. If only one service location 120 has the greatest probability or a sufficient probability of servicing the work within the target time, that one service location 120 is selected (step 312). If multiple service locations have been determined to have the greatest probability of servicing the work within the target time, (*i.e.* if the greatest probability is calculated with respect to multiple service locations), or if multiple service locations have a sufficient probability of servicing the work within the target time, the service location 120 having the most favorable advance time metric is selected from the multiple service locations 120 having the greatest or sufficient probability of servicing the work within the target time (step 316). At step 320, the work is assigned to the service location 120 selected at step 312 (if only one service location 120 has the greatest probability or a sufficient probability of servicing the work within the target time) or to the service location 120 selected at step 316 as having the most favorable advance time metric (if multiple service locations 120 were determined to have a greatest probability or a sufficient probability of servicing the work within the target time). The process of assigning a work request then ends (step 324), at least until a next service request is received or generated.

With reference now to **Fig. 4**, the determination of the probability that a service location 120 will be able to service work within a target time relative to other service locations 120 in accordance with an embodiment of the present invention is illustrated. Initially, at step 400, the estimated wait time (EWT) for a selected service location 120 is calculated. The estimated wait time may be calculated using various methods known to the art. For example, the estimated wait time may be calculated by determining an average rate of advance for a service location 120, and in particular for a queue established in connection with a service location 120, by multiplying the average rate of advance by the position of the next work request to be received, as described in U.S. Patent No. 5,506,898, the disclosure of which is incorporated herein by reference in its entirety.

At step 404, a determination is made as to whether the estimated wait time is greater than the target service time that has been established. If the estimated wait time at the service location 120 exceeds the target service time, the number of opportunities for servicing a work request within the target time (#OPPS) is set to zero (step 408). If the estimated wait time is not greater than the target service time, the weighted advance time (WAT) for the queue associated with the service location 120 is calculated (step 412). The weighted advance time is the measure of the average time that is required for a work request to advance one position in the queue. Accordingly, the weighted advance time may be calculated as a continuously updated average advance time. As can be appreciated by one of ordinary skill in the art, the time period over which advance times are averaged for a queue can be varied.

At step 416, the number of opportunities for work to be serviced within the target time is calculated. In accordance with an embodiment of the present invention, the calculation of opportunities for work to be serviced within the target time is calculated using the algorithm: #OPPS = ((Target time-EWT)/WAT)+1, where Target time is the target time for servicing the work. The number of opportunities for the queue associated with the service location 120 set or determined at step 408 or step 416 is then recorded (step 420).

After recording the calculated number of opportunities for the service location 120, a determination is made as to whether queues associated with additional service locations 120 are applicable to the work request (*i.e.* are eligible) (step 424). If additional service locations 120 are available, the next service location is gotten (step 428) and the system returns to step 400. If additional service locations are not available, the service location or locations 120 having the greatest number of opportunities to service the work request, or the location or locations 120 having a sufficient probability of servicing the work request, are set equal to the location or locations 120 having the greatest probability (or sufficient probability) of servicing the work request within the target time (step 432). In accordance with an embodiment of the present invention, a service location 120 having a sufficient probability may be identified by comparing a calculated number of opportunities for that service location 120 to a preselected number of opportunities deemed to correspond to a sufficient probability. The process for determining the relative probabilities of service locations 120 then ends (step 436).

The method generally set forth in connection with the flow chart shown in **Fig. 4** is suitable for use in connection with step 304 of **Fig. 3**.

With reference now to **Fig. 5**, the calculation of an advance time metric in accordance with an embodiment of the present invention is illustrated. In particular, **Fig. 5** illustrates a method for calculating an advance time metric comprising a weighted advance time trend, and can be used to select a single service location 120 from a number of service locations 120 in connection with step 316 of **Fig. 3**. Initially, at step 500, the weighted advance time for a service location 120 is calculated. In general, the calculation of the weighted advance time for a particular service location 120 will have been performed as part of determining the relative probability that the service location 120 will complete a work request within the target time. Accordingly, the WAT may be received at step 500. At step 504, the WAT change is calculated. The WAT change may be calculated as: $WAT\_Change = (WAT_n - WAT_{n-1})/ WAT_{n-1}$. For example, if at time 'n-1' WAT=10, and then at time 'n' WAT=9, WAT_Change = (9-10)/10 = -0.1. A negative number means that WAT is trending downwards, by a ratio of 0.1 in this case. That is, the WAT has become 10% smaller. At step 508, the WAT trend is calculated. The WAT trend is an exponential moving average of the WAT changes. The WAT trend may be calculated as $WAT\_Trend_n = (x*WAT\_Trend_{n-1}) + ((1-x)*WAT\_Change)$ where x is a constant such as 0.9. In other words, WAT_Trend is an exponential moving average, which determines if WAT is trending downward or upwards and at what rate. If WAT is trending downwards, this is a positive sign that conditions may be improving for this service location 120. All other things being equal, a service location 120 that is showing the best signs of improvement is preferred. Next, the calculated WAT_Trend for the service location 120 is recorded (step 512). At step 516, a determination is made as to whether additional service locations 120 are available. For example, a determination of

whether an additional service location having a greatest or sufficient probability of completing work within the target time is available may be made. If an additional service location 120 is available, the system gets the next service location 120 (step 520) and returns to step 500. If an additional service location 120 is not available, the service location 120 having the lowest calculated WAT_Trend is set equal to the service location 120 having the most favorable advance time metric (step 524). The process for determining an advance time metric then ends (step 528).

In accordance with another embodiment of the present invention, the advance time metric used to select one of a number of service locations 120 having a greatest probability, or a sufficient probability, for servicing the work within the target service time at step 316 of **Fig. 3** is the estimated wait time associated with each service location. In particular, the work is assigned to the service location 120 included among the service locations 120 determined to have the greatest or a sufficient probability with the lowest estimated wait time. According to such an embodiment, at step 316 of **Fig. 3**, the service location 120 having the lowest expected wait time is selected from the service locations 120 having the greatest or a sufficient probability of servicing the work within the target time.

As can be appreciated from the foregoing description, multiple service locations 120 may be determined to have a greatest probability of servicing work within a target time period if more than one service location 120 is determined to have the highest calculated probability. Thus, in connection with embodiments of the present invention in which relative probability is calculated as a number of opportunities to complete work within a target time period, multiple service locations 120 have the highest probability if

they have the same highest number of opportunities. For example, if a first service location is determined to have three opportunities, a second service location 120 is also determined to have three opportunities, and a third and final service location 120 is determined to have two opportunities, the first and second service locations 120 each have the same greatest probability of servicing the work within the target time.

As can also be appreciated from the foregoing description, multiple service locations 120 may be determined to have a sufficient probability of servicing work within a target time if the calculated number of opportunities exceeds a number preselected as being sufficient. For example, if three opportunities to service work within a target time is selected as representing a sufficient probability that the work will be serviced within the target time, and a first service location 120 is determined to have four opportunities, a second service location 120 is determined to have three opportunities, and a third and final service location 120 is determined to have two opportunities, the first and second service locations 120 both have a sufficient probability of servicing the work within the target time.

The foregoing discussion of the invention has been presented for purposes of illustration and description. Further, the description is not intended to limit the invention to the form disclosed herein. Consequently, variations and modifications commensurate with the above teachings, within the skill and knowledge of the relevant art, are within the scope of the present invention. The embodiments described hereinabove are further intended to explain the best mode presently known of practicing the invention and to enable others skilled in the art to utilize the invention in such or in other embodiments and with various modifications required by their particular application or use of the

invention. It is intended that the appended claims be construed to include the alternative embodiments to the extent permitted by the prior art.